RL

SEP 11 1984

Final Report PFTR-1092-83-1
Contract Number: N00014-80-C-0150
Work Unit Number: NR197-064

1

THE PSYCHOLOGY OF CONFIDENCE

FINAL REPORT

Prepared for

OFFICE OF NAVAL RESEARCH
800 North Quincy Street
Arlington, Virginia 22217

PERCEPTRONICS
6271 VARIEL AVENUE • WOODLAND HILLS • CALIFORNIA 91367 • PHONE (213) 884-7470

87 1

SUMMARY

This report is the final report of research done under the contract THE PSYCHOLOGY OF CONFIDENCE (Contract N00014-80-C-0150) awarded by the Office of Naval Research to Perceptronics, Inc. The period of the contract was from January 15, 1980, to January 15, 1983.

The report is divided into two sections. The first section is a narrative summary of the research completed under this contract. It describes the general background to our work on confidence and outlines the strategy used in this research program. In addition to describing the characteristics of a good probability assessor, it gives an overview of the quality of people's performance as probability assessors and offers an interpretation of our research results that may be useful in improving the appropriateness of confidence judgments.

The second part of this report is a compendium of the technical reports and archival publications produced under this contract. Each report is listed according to author, title and technical report reference. Archival references are provided where appropriate. Included with each technical report reference is an executive summary of the body of the report.

Introduction

All knowledge has two components, a substantive statement about some aspect of the world and an evaluative indication of the validity of that statement. The substantive statement is one's best guess regarding a state of the world; the evaluative component, an expression of confidence, captures how good that best guess appears to be. In making decisions, one must use both components of knowledge. Depending on one's level of confidence, one may decide to take firm action, hedge one's bets, collect more information, or just think harder. The wisdom of decisions is constrained by both the extent of one's substantive knowledge and the appropriateness of one's confidence. All other things being equal, we should be better off the more we know. However, if we underestimate our knowledge, we may be unable to capitalize on it; if we overestimate our knowledge, we may act precipitously.

Although the importance of confidence as a controlling factor both in processing information and in governing subsequent behavior is widely acknowledged, the details of its involvement have not been carefully elaborated within cognitive psychology. For example, although the "cognitive revolution" has highlighted the importance of control processes in memory search, confidence is seldom measured or given any explicit representation in theoretical models.

Confidence has been studied most extensively in the context in which its role is most obvious, decision theory. Decision theory assumes that probabilities express individuals' degrees of belief. Thus, when one states the probability associated with a particular state of the world, one is expressing one's confidence in one's knowledge. The experimental study of confidence as expressed via probabilities has gradually produced both a viable research methodology and some robust findings.

One recurrent finding is that people are often overconfident; they have unwarranted or exaggerated faith in the correctness of their knowledge. A second finding is that this overconfidence is related to the difficulty of the task; the more difficult the task, the greater the discrepancy between people's accuracy and their confidence.

Overconfidence has been documented with a large variety of tasks and subject populations. However, experiments have shown that overconfidence can be reduced by training. In addition, overconfidence has been lessened by requiring people to make explicit lists of possible reasons why they might be wrong. This finding suggests that one reason for overconfidence is people's tendency to rely primarily on reasons why their view of facts are correct without giving attention to reasons why they might be wrong.

Previous research (more thoroughly reviewed in the previous proposal for this project) has provided considerable evidence on some of the concommitants of confidence; however, it has been incomplete. One goal of this project was to fill the gaps in our knowledge about what kinds of people, items, procedures, and contexts lead to overconfidence. A further goal was to explore methods for correcting overconfidence in ways that may shed light on the cognitive processes that produce it.

We hoped, in the course of this project, to develop a theory of confidence. Our initial framework for understanding confidence was a series of sequential steps that, logically, one would go through to evaluate one's confidence in the answer to a factual question:

(a) Read the question.

(b) Consider the context in which it arose.

(c) Search memory for relevant information.

(d) Refer to relevant non-memory sources (e.g., books, experts).

(e) Arrive at one's best guess at the correct answer.

(f) Weigh the validity and completeness of relevant evidence supporting and contradicting that answer.

(g) Translate the resultant feelings of confidence into a verbal or numerical expression of confidence.

We attempted to explore the psychological processes and the pitfalls involved in each of these steps, to see where and why overconfidence emerges. In this effort, we tried to strengthen the links between the literature and concepts of decision theory and those of cognitive psychology.

Research

Reviews

The project began with two comprehensive reviews of the literature on confidence and overconfidence, each from a somewhat different perspective. Between them, they set the stage for the empirical work that followed. Lichtenstein, Fischhoff and Phillips (1982) reviewed all published (and some unpublished) studies that evaluated the validity of confidence assessments using a measure called calibration. A set of assessments is considered to be well calibrated if XX% of the assessments of .XX are associated with correct answers. For example, if the assessors are correct 70% of the time that they are .70 confident.

On the methodological side, the review revealed the difficulty, or subtlety, of studying confidence assessments, many of which can imperil the interpretion of results. One example is the need for large samples of data in order to produce stable calibration curves; without stability the picture of performance that is derived could mislead an investigator studying someone else or give unreliable feedback to assessors trying to improve their own performance. A second example is the availability of different statistical measures for summarizing assessment performance, each suitable for somewhat different purposes. A third example is the fact that calibration is highly dependent upon the difficulty of the questions being addressed. Thus, substantial errors can be (and have been) made as a result of comparing performance obtained in situations wherein the assessors had different levels of knowledge.

Substantively, the review revealed overconfidence to be the predominant pattern in people's assessments. It traced that overconfidence to a relative insensitivity to the extent of one's knowledge. In general, as individuals become more knowledgeable they also become more confident. However, the increase in confidence far outstrips the increase in knowledge; for example, as confidence increases from .5 to 1.0, the percentage of correct answers might increase from 50% to 80%. Typical studies, like many typical real-life situations, present items of moderate to great difficulty, as there is often less interest in confidence regarding easy questions. At these difficulty levels, the predominant pattern is overconfidence. A typical summary might be a mean confidence of .80 associated with 65% correct responses. The worst performance was associated with the highest expressions of confidence, such as being 80% correct when 100% confident. Decisions relying on assessments of uncertainty may be particularly sensitive to such overconfidence at the extremes. Figure 1 shows a set of typical empirical results.

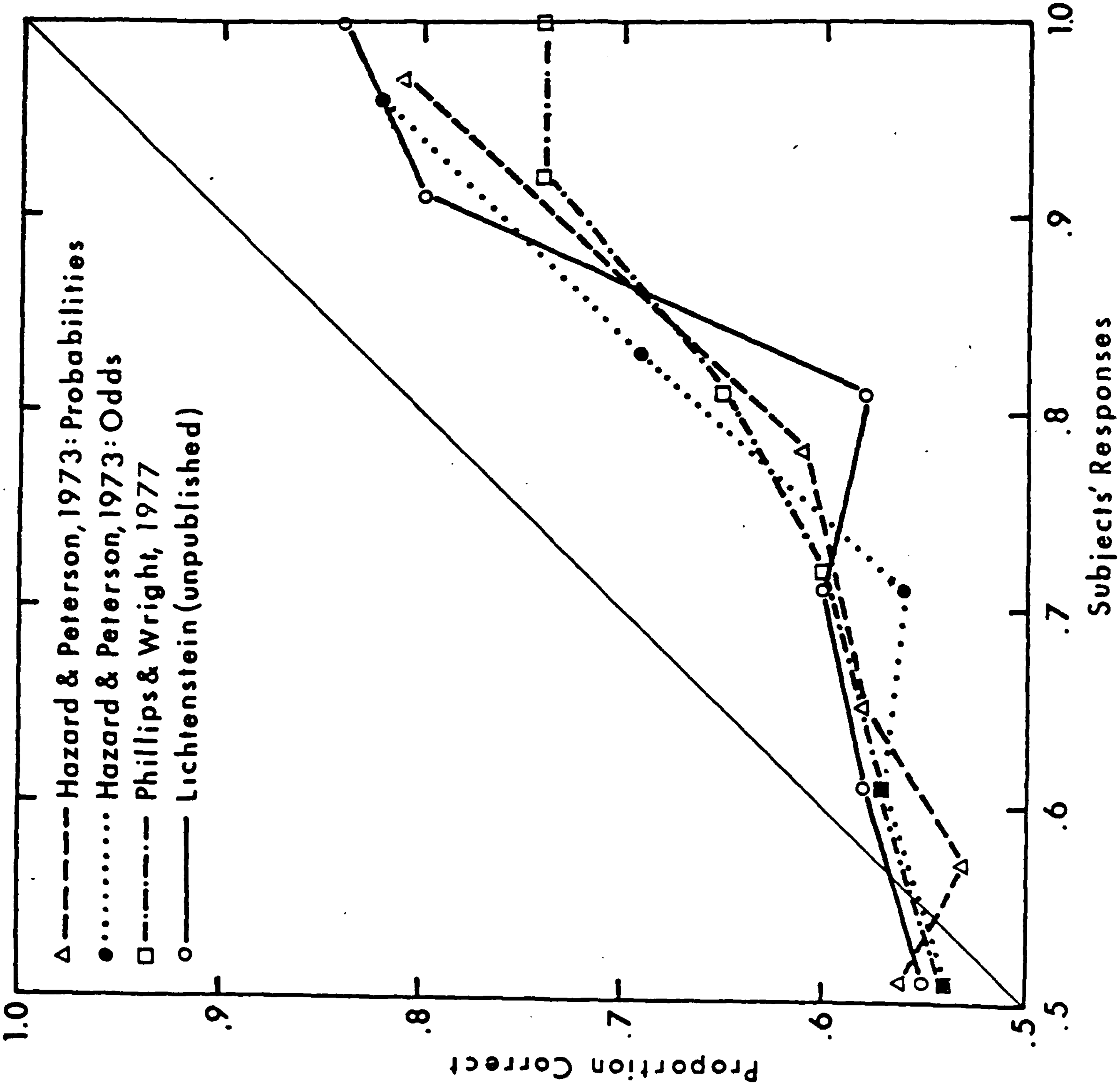
△ Hazard & Peterson, 1973: Probabilities
● Hazard & Peterson, 1973: Odds
□ Phillips & Wright, 1977
○ Lichtenstein (unpublished)
Proportion Correct
1.0
.9
.8
.7
.6
.5
Subjects' Responses
.5
.6
.7
.8
.9
1.0

Figure 1

Like most of the studies upon which it was based, this review was stronger on statistical theory than on psychological theory. A second review of the same material (Fischhoff, 1982) emphasized the psychological dimension as well as the applied implication of having to cope with overconfidence. Specifically, it developed a "theory of debiasing," a framework for ways in which one could attempt to eliminate bias in any form of judgment, including confidence assessments. The framework is based upon how the bias is attributed, whether to the judge, the task, or the match between the judge and the task. The first of these attributions assumes that people lack the cognitive skills needed to perform the task. The second assumes that the task is somehow unfair to the judges, and prevents them from performing to the best of their ability. The third attaches blame to neither, rather it assumes that the natural (and fair) presentation of a task typically leads to sub-optimal performance; however, it is possible to restructure the task so that it evokes more of respondents' intellectual capabilities. This is the most psychologically rich attribution for it forces the investigator to consider how people think and how they might be helped to think better.

Within each of these categories there are a number of subcategories all of which are presented in Table 1. The full set of published studies on overconfidence was reanalyzed in terms of this scheme. It revealed some stable and informative patterns. One was that attempts to eliminate the bias by eliminating conceivable methodological artifacts had failed. It made no difference, for example, if one raised the stakes further, clarified the instructions or varied the homogeniety of the questions. What did seem to make a difference were some ways of changing the judge and changing the interaction between judge and task. Intensive training programs will prompt

Table 1

Debiasing Methods

According to Underlying Assumption

Assumption	Strategies
Faulty tasks	
Unfair tasks	raise stakes clarify instructions/stimuli discourage second guessing use better response modes ask fewer questions
Misunderstood tasks	demonstrate alternative goal demonstrate semantic disagreement demonstrate impossibility of task demonstrate overlooked distinction
Faulty judges	
Perfectible individuals	warn of problem describe problem provide personalized feedback train extensively
Incorrigible individuals	replace them recalibrate their responses plan on error
Mismatch between judges and task	
Restructuring	Make knowledge explicit search for discrepant information decompose problem consider alternative situations offer alternative formulations
Education	rely on substantive experts educate from childhood

personalized feedback and improved performance both in laboratory studies and in field experience with weather forecasters. Simply warning people against overconfidence (or exhorting them to be careful) does not have this effect. In the area of changing the structure of tasks (without providing training or information), the most effective manipulation was requiring judges to give reasons why the answer that seemed most likely to be correct might have been wrong. This requirement seemed to increase the portion of correct answers (perhaps by helping people to detect errors) and decrease confidence (perhaps by helping judges to recall information that they had but did not spontaneously retrieve).

The same kind of secondary analysis was also performed on the full set of published studies concerning another judgmental bias, hindsight bias, the tendency to exaggerate in hindsight what could have been known in foresight. Hindsight bias is a likely contributor to overconfidence. An exaggerated feeling that one knew all along what would happen could heighten one's feeling of knowledgeability, or of having nothing to learn. Although the evidentiary record here was sparser, the pattern of results was quite similar. Exhortation and technical changes in the task and exhortion made no difference. However, the bias was reduced by forcing respondents to list reasons why they might be wrong.

Both reviews also suggested that mere substantive expertise in an area was not a guarantee of improved performance. There seemed to be a need for it to be accompanied by some specific training or procedures directed at the improvement of judgment (and not just the acquisition of facts).

These reviews conveyed a picture of the strengths and weaknesses in the research literature, and that they directed the design and conduct of the empirical studies that followed. Those studies then guided the theoretical analysis that either summarized the research or elaborated local points.

Empirical Research

Our initial empirical work was designed to clear up a number of residual methodological points that emerged from the reviews. One looked at the question of whether any possible misunderstanding of the task might have degraded subjects' performance. Subjects here (Lichtenstein & Fischhoff, 1981) were given extensive instructions in the meaning of confidence judgments, with particular attention being paid to explication of the calibration measure that would be used to evaluate subjects' performance. This manipulation had no effect on any observable aspect of subjects' performance. Its results demonstrated the robustness of the overconfidence phenomenon and increased our faith in previous research using simple instructions, and legitimated using simpler instructions in future research.

A second methodologically oriented study introduced a marked difference in how subjects provided their assessments. Instead of having them consider each of a series of items sequentially, this study (Fischhoff, MacGregor & Lichtenstein, 1982) had them first sort the items into piles of items about which they were equally knowledgeable. After repeated sorting, they then were asked to assess the probability of being correct for the typical item in the pile. Different experimental groups were instructed to use different numbers of piles. This procedure differs from the standard one in a number of ways that might be psychologically significant. It places the emphasis on evaluating knowledge rather than on the production of a numerical response expressing that perceived level of knowledge. It allows comparison between items. It allows an assessment of any interdependence among items. It allows an assessment of the overall difficulty of the item pool. It allows a second and third look at questions, hence, further opportunity to consider the best answer and appropriate confidence in it. Nonetheless, there was no difference in performance, compared with the typical response mode. Had there been differences, then we would have had some building blocks for a richer substantive theory of confidence assessment. As it was, though, we had further evidence of the robustness of overconfidence and a study that, in retrospect, seemed primarily methodological in character.

In order to evaluate calibration, the investigator needs to be able to score the judge's answers as correct or incorrect. Frequently, the easiest way to accomplish this is frequently to take archival data, regarding which the judge has some knowledge on the basis of which to infer the correct answer. Questions fitting this niche might refer to last year's economic parameters, facts from textbooks, or items gleaned from an almanac. One might wonder whether the fact that the answers are known to someone does not in some way affect judges' responses to them. A priori, one might speculate that it increases expressed confidence, because one feels that one should know about such things, or that it decreases expressed confidence because one is concerned about being evaluated by someone who will readily detect exaggeration. A way to avoid this problem is to collect assessments about future events, regarding which neither subject nor investigator knows the answers. A study by other investigators seemed to reveal a difference between responses to past and future events; however, it confounded differences in the temporal setting with differences in item difficulty. Subjects were less overconfident regarding the future events; however, those events were also easier than the contrasted past events, a difference that one would predict to produce such a difference in the level of overconfidence.

We conducted a series of three experiments in which subjects first predicted the outcome of a variety of events that would be resolved in the following month (e.g., the results of election or sporting competitions) and then indicated their confidence in those predictions (Fischhoff & MacGregor, 1982). Once the results were known we were able to evaluate the difficulty of the predictions. Then, we relied upon our extensive collection of calibration data to produce a previously collected set of responses to past items of comparable difficulties. We found almost no difference whatsoever in response patterns or performance. Confidence in predictions seemed to be produced by the same processes as confidence in items of past knowledge. Again, had we found any differences, they would have helped enrich our theory of those processes. As it was, they strengthened previous notions and gave this study a methodological cast.

The one modest but interesting peripheral result was that with future events we found a somewhat higher proportion of subjects who never used the extreme response of 1.0, expressing complete confidence. Moreover, these subjects were somewhat less overconfident beyond any differences attributable to differences in knowledge level. Although not a large effect, the spontaneous non-use of 1.0 suggests a weak individual differences variable in probability assessment. Search for this effect in other experiments in this project produced a mixed pattern. Non-users were never more overconfident and sometimes less than users of the 1.0 response.

Several other studies looked more intensively at characteristics of judges that might give some clues as to their thought processes. One examined the difference between the response patterns of men and women (Lichtenstein & Fischhoff, 1981). A common observation regarding sex differences in our society is that men are socialized to be confident, whereas women are trained to be modest, or even deprecatory, about their abilities. If this were the case, then less overconfidence might be expected of women than of men. If that were the case, then one could capitalize on the research literature on sex differences for insight into the cognitive processes involved. To that end, we compared groups of men and women of comparable education level on a large set of general knowledge items. The women were slightly less overconfident; however, that difference was entirely attributable to their having slightly less knowledge regarding the answers to this particular set of items. Here, too, we had further evidence of the robustness of overconfidence.

Another dimension of individual differences produced quite dramatic effects (Lichtenstein & Fischhoff, 1980). We submitted a set of 500 general knowledge items to 8 academic experts in confidence assessments, 15 individuals who had one year previously received extensive training in probability assessment, and 12 untrained individuals. These untrained individuals were, like those observed elsewhere, relatively insensitive to the extent of their own knowledge and typically overconfident for all but the easiest of questions. The previously trained individuals apparently had preserved some of their acquired abilities and showed somewhat superior performance. The academic experts, all of whom were familiar with the literature on overconfidence, generally eliminated that bias in their aggregate responses. Indeed, on the average, their mean confidence was somewhat less than their mean proportion of correct answers. However, they were not any more sensitive to differences in their level of knowledge. That is, their calibration curves were just as flat as those of untrained individuals. What was different was their mean level of confidence. We believe that this was a technical correction and that such a correction is all that can be derived on the basis of reading the research literature. Gaining greater sensitivity requires changes in how people think, not just in the kinds of numbers that they produce. The fact that these experts did change their mean responses is evidence of people responding to aggregate results regarding other people's performance; such learning from others experience has not been common.

Our work on investigating the way in which the match between task demands and respondents' cognitive skills focussed on two topics. One of these was prompted by the result revealed by the reviews, to the effect that performance improves when people are required to think of reasons why they might be wrong. A number of possible reasons for this effect come to mind, one of which is artifactual. It holds that the repeated reminder to "think of why you might be wrong" carries with it a message to reduce one's expressed confidence, whatever one's subjective feeling is. The worry is diminished by the fact that previous research has shown respondents to be insensitive to communications regarding the overconfidence of others like them, as well as to explicit instructions "to feel free to respond with .5 to all of a set of [two-alternative] questions if you are just guessing at the answer to each." To reduce any residual concern, Beyth-Marom and Lichtenstein (1983) used a procedure that eliminated any fear of such implicit communication to subjects. Specifically they used questions for which only a single possible alternative was presented. Responses to these questions were then compared to responses to versions of the same questions in which a second possible answer was appended. Provision of a second possible answer should prompt at least some search for contradictory reasons without suggesting that subjects doubt their own abilities. It succeeded in reducing subjects' overconfidence and improving their calibration relative to the one-alternative group. It was not, however, enough to eliminate all problems (being, in fact, the standard condition in which those effects had previously been observed). This set of data also produced mixed results on two previously observed effects. Consistent with a previous finding, spontaneous non-users of 1.0 were less overconfident. However, when subjects in a separate condition were asked to provide reasons for and against their answers, performance did not improve.

A series of five experiments by Beyth-Marom and Fischhoff (in press) explored further the way in which people treat supporting and contradicting evidence when evaluating a possible answer. It showed that they often have a poor appreciation of the role of competing hypotheses when evaluating the support that a piece of evidence confers on a focal hypothesis. When asked to test the validity of an hypothesis, H, in the light of a datum D, only half of subjects expressed an interest in P(D/$\underline{H}$), an essential piece of information. The main subjective cue to the validity of H was the magnitude of P(D/H), even though a high P(D/H) is neither necessary nor sufficient for D to increase confidence in H. Further experimentation showed that this bias was reduced by explicitly asking subjects to evaluate whether H or $\underline{H}$ was true (and not just to evaluate the validity of H). This can be compared to the difference between the one- and two-alternative groups in Beyth-Marom and Lichtenstein (1983). When subjects were presented with both P(D/H) and P(D/$\underline{H}$), they revealed a qualitative understanding of their interrelationship. What seems to be missing is the spontaneous realization of the relevance of competing alternatives. The fact that people can make use of competing alternatives when presented, just as they can produce and use contradictory information once prompted to do so, suggests that something can be done to improve the match between tasks and the cognitive skills that people apply to them.

In many situations, where people must choose among competing hypotheses, the observable response is not a probability assessment, but a categorical choice. In making those choices, it is common to have both background (or a priori) information regarding the general likelihood of the respective categories and case-specific information pertinent to the particular prediction. Much previous research has indicated that case-specific information has an undue influence on people's categorical choice. A theoretical analysis by Beyth-Marom and Fischhoff (1981), which is described in the following section, suggested that patterns of categorical choices can be clarified by more detailed consideration of the confidence that people place in them. In a study of hypotheses generated by this theoretical analysis, Fischhoff and Beyth-Marom (1981) examined categorical predictions in situations wherein the amount of justifiable confidence was varied systematically. This was done by keeping the data constant, but varying the differentiability of the competing hypotheses. The results showed that people relied on a judgmental heuristic called "representativeness" (Kahneman, Slovic & Tversky, 1982) whenever that was at all possible. Although validation of these responses was not possible here as it had been in the confidence studies, circumstantial evidence indicated that the ability to rely on a plausible rule (representativeness resulted in confidence levels that were hard to justify). There was, however, a reduction in confidence judgments when subjects made predictions contrary to the background information. Thus, the confidence judgments revealed a (modest) level of sophistication that could not be seen in the categorical predictions.

Theoretical Work

Our initial intent was to flesh out a full cognitive theory of confidence assessment according to the seven-stage process presented in the introduction to this summary. We still believe that that is an insightful perspective; indeed, it is the one that prompted the empirical studies that we have conducted. However, it seems premature to present such a theory. There are far too many gaps in our knowledge to fill in the pieces adequately. Moreover, the deeper we probed into the theoretical basis of confidence judgments, the more we realized the need to clarify some of the accepted fundamentals. As a result, our theoretical efforts have taken a somewhat different turn. In addition to the theoretical components of the literature reviews, secondary analyses, and empirical papers, we have produced two analyses of specific topics in probability assessment and a comprehensive treatment which recently was published in *Psychological Review*.

The first of these theoretical analyses considered the conditions under which one would be expected to be well calibrated (Kadane and Lichtenstein, 1982). It starts from the assumption that one is a coherent Bayesian in the sense of having a set of beliefs that obey the probability axioms. Under this (fairly rigorous) assumption, perfect calibration is shown formally to be expected either when one has accurate feedback regarding previous predictions or when the future events that are to be predicted can be viewed as independent. These results point to the need for empirical studies of how performance is affected when these conditions are not fully met.

The second specific analysis considered the issue in psychological (as opposed to statistical) theory of how people integrate case-specific and background information into an overall appraisal of the probability of an hypothesis' validity. It was prompted by an article claiming to demonstrate that, contrary to previous results, people can be very attentive to background information even in the presence of seemingly useful case-specific information. Beyth-Marom and Fischhoff (1981) argued that a deeper analysis of the logical and psychological character of tasks used in the various studies, allows one to reconcile the seemingly conflicting evidence observed in them. They offer a simple account for existing data, namely that people will rely primarily on individuating evidence unless it fails to provide a guide to prediction. However, they do reduce their confidence when making predictions contrary to the background information, particularly when they make a series of predictions offering performance feedback. Whether those changes in confidence and effects will be observed depends upon whether the opportunity is given for their expression in this light. The research on confidence allows a deeper understanding of how confidence is translated into action, whereas the research on categorical predictions provides insight into the formation of confidence.

In performing these studies, we repeatedly found ourselves referring back to the logical framework provided by the Bayesian scheme for hypothesis evaluation. It is that scheme which gives a central role to subjective feelings of confidence and which guides the translation of those beliefs into optimal actions. It provides a standard for judging both the internal validity (or coherence) of a set of beliefs and the adequacy of their external validity (or calibration). Moreover, it has provided the touchstone for many of the studies of cognitive processes and performance upon which our research was based. However, in reviewing the relevant literature we felt that there was no adequate comprehensive exposition of normative Bayesian inference or of the attempts to compare intuitive performance with it. We decided to develop such an exposition, both as a service to the field and as a vehicle describing and integrating our own work (Fischhoff & Beyth-Marom, 1983).

The resulting essay begins by presenting the basic Bayesian model of inference. It then exploits its descriptive potential by identifying a set of logically possible forms of Bayesian behavior. A review of empirical studies demonstrates which of these potential biases have, in fact, been observed. That review also shows that the Bayesian interpretation of individual studies requires consideration of how all aspects of the Bayesian model are addressed in them. In particular, it shows that some results in the literature that are presented as biases may not be biases at all, that sometimes two apparently different biases are special cases of the same deviation from the model, and that in some cases the same description has been given to rather different phenomena (e.g., the confirmation bias). All in all, it offers the promise of greater simplicity in accounting for empirical results at the expense of requiring richer interpretions of each.

Implications

We believe that this project leaves us with a considerable sense of closure regarding some aspects of the psychology of confidence. The methodology of studying confidence has been advanced to a point where future investigators can design experiments with clear guidance as to what methods to use and what pitfalls to avoid. The literature on confidence has been comprehensively summarized from several different perspectives, suitable for different purposes. Those summaries have revealed some very robust patterns of results, some of which approach the status of "scientific facts," in particular the finding of insensitivity to degree of knowledge expressing itself in overconfidence. Our empirical studies have also shed light on the factors that do and, more frequently, the factors that do not affect people's confidence-assessment process. One of our theoretical pieces integrates the confidence assessment work with a large body of other judgment research within the context of Bayesian inference. Another clarifies the relation between orderly beliefs and appropriate levels of confidence. A third simplifies the empirical results regarding the relationship between confidence and action. We believe that together with previous results we have a robust set of conclusions and a sound research basis for future studies.

Regarding the elicitation and utilization of confidence assessments in applied settings, the research provides several kinds of useful knowledge. First, it shows the degrees of freedom that one has in structuring elicitation sessions. Specifically, the robustness of the assessments indicates that one has considerable freedom in adapting the elicitation procedure to the needs and desires of the judge. Secondly, the research gives a solid basis for anticipating the quality of the probability assessments likely to be observed in various circumstances. For those who must rely on those assessments, these results can show how seriously to take them and what corrections might be in order. Finally, the research shows what procedures are most effective (and what procedures are totally ineffective) for improving the quality of confidence assessments. As explicit probability judgments increasingly become a part of decision making and policy setting procedures, such knowledge is essential to regulating and exploiting them.

Summaries of Technical Reports and Archival Publications

Produced under ONR Contract N00014-80-C-0150

THE PSYCHOLOGY OF CONFIDENCE

to

Perceptronics, Inc.

Title: Lichtenstein, S. & Fischhoff, B. How well do probability experts assess probabilities? (Tech Rep. PTR-1092-80-8). Perceptronics, Inc., August 1980.

Summary

Past research on people's ability to assess probabilities has shown two common errors, overconfidence in one's knowledge and insensitivity to task difficulty. This research has created a new class of experts: those who have studied probability assessors and who are aware of the common errors. The performance of eight such experts is here compared to the performance of 12 untrained subjects and 15 who had previously received training in probability assessment. All subjects responded to 500 general-knowledge items whose difficulty could be measured *a priori* from the item content. The experts appeared to have overcorrected for the overconfidence error: they were notably underconfident, whereas the untrained subjects were overconfident and the trained subjects were mixed. The experts were more sensitive than the other two groups to variations in item difficulty. However, even they showed a substantial insensitivity to difficulty, relative to ideal performance. Introspection suggests that this second error would be hard to overcome.

Title: Fischhoff, B. Debiasing (Tech. Rep. PTR-1092-81-3). Perceptronics, Inc., March 1981.

Archival Publication: Fischhoff, B. Debiasing. In D. Kahneman, P. Slovic and A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, 1982.

Summary

Research over the last decade has identified a number of powerful biases in people's judgments. Because of the threat that those biases pose to judgments and the decisions based on them, a focus of research has been ways to eliminate them. The present paper begins by developing a conceptual framework for characterizing debiasing procedures in terms of their underlying theory of psychological processes. It then applies this framework to published studies on two judgmental biases: hindsight bias, the tendency to exaggerate in hindsight how much we (or anyone else) could have known in foresight, and overconfidence, the tendency to exaggerate how much one knows.

From this reanalysis, a surprisingly coherent picture seems to emerge. Neither bias is appreciably reduced by simply exhorting people to work harder, raising the stakes riding on their judgments, or other "mechanical" manipulations. Nor do substantive experts dealing with subject matter in their areas of expertise seem to have any particular immunity—unless they have received training in judgment per se. Such training also has proven capable of improving the performance of non-experts. The only other manipulation that has proven effective involves restructuring the tasks so as to be more compatible with the judge's natural thought processes.

Although this picture is not complete (i.e., some studies remain to be done; similar reviews are needed for efforts to eliminate other biases), it does allow some tentative conclusions for theory and practice. One is that these biases are not just artifacts of experimental procedure, but seem

endemic to intuitive judgment. A second is that either training programs or other judgmental aids can be useful in reducing biases. The third is that those aids are most efficiently developed on the basis of an understanding of the cognitive processes that produce biases.

Title: Bar-Hillel, M. & Fischhoff, B. When do base rates affect predictions? (Tech. Rep. PTR-1092-81-5). Perceptronics, Inc., May 1981.

Archival Publication: Bar-Hillel, M. & Fischhoff, B. When do base rates affect predictions? Journal of Personality and Social Psychology, 1981, 41, 671-680.

Summary

One of the most fundamental and recurrent judgment tasks involves combining base-rate information and individuating information into a prediction. The former tells what typically happens in such situations; the latter tells something special about the particular case in question. According to statistical principles, the relative importance one attached to the two kinds of information should depend upon the relative quality of the evidence each provides.

Empirical studies have shown, however, that people tend to ignore base rates entirely in the presence of even the most flimsy of individuating evidence. Such a tendency would threaten the validity of many judgments and indicate the need for either decision aids or training. As a result, many studies have tried to circumscribe the range of this "base-rate effect." One recent study concluded (a) that base rates are ignored when they are presented in summary (e.g., X% of the time, Y happens"), but not when they are presented as a series of cases; and (b) that the effect is reduced when judges make a series of predictions rather than just one.

The present paper begins with a reanalysis of the data and experimental design of that study, showing that neither of these conclusions follows necessarily from the evidence presented there. In order to do so, it offers an alternative framework for thinking about and studying base-rate phenomena. It concludes by arguing for a more parsimonious account of existing data on this question: People will rely primarily on individuating evidence unless it

fails to provide a guide to prediction. However, people may be less confident when making predictions contrary to the base rate, particularly when they are making a series of predictions that offer some feedback on their performance. Whether that reduced confidence is evident will depend upon whether the task offers an opportunity for its expression. From this perspective, the evidence accumulated to date allows one to make fairly precise predictions of when base rates will affect predictions.

Title: Lichtenstein, S., Fischhoff, B. & Phillips, L. D. Calibration of probabilities: State of the art to 1980 (PTR-1092-81-6). Perceptronics, Inc., June 1981.

Archival Publication: Lichtenstein, S., Fischhoff, B. & Phillips, L. D. Calibration of probabilities: State of the art to 1980. In D. Kahneman, P. Slovic and A. Tversky (Eds.), Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, 1982.

Summary

This paper presents a comprehensive review of the research literature on an aspect of probability assessment called "calibration." Calibration measures the validity of probability assessments. Being well-calibrated is critical for optimal decision-making and for the development of decision-aiding techniques.

Subjective probability assessments play a key role in decision making. It is often necessary to rely on an expert to assess the probability of some future event. How good are such assessments? One important aspect of their quality is called calibration. Formally, an assessor is calibrated if, over the long run, for all statements assigned a given probability (e.g., the probability is .65 that "Romania will maintain its current relation with People's Republic of China for the next six months."), the proportion that is true is equal to the probability assigned. For example, if you are well calibrated, then across all the many occasions that you assign a probability of .8, in the long run 80% of them should turn out to be true. If, instead, only 70% are true, you are not well calibrated, you are overconfident. If 95% of them are true, you are underconfident.

While this characteristic of assessors has obvious importance for applied situations, people's calibration has rarely been discussed by decision analysts or decision advisors. In the last few years, there has developed an extensive literature about calibration, reporting both laboratory and real-

world experiments. It is now time to review this literature, to look for common findings which can be used to improve decisions, and to identify unsolved problems.

Findings

Two general classes of calibration problem have been studied. The first class is calibration for events for which the outcome is discrete. These include probabilities assigned to statements like "I know the answer to that question," "They are planning an attack," or "Our alarm system is foolproof." For such tasks, the following generalizations are justified by the research:

1. Weather forecasters, who typically have had several years of experience in assessing probabilities, are quite well calibrated.

2. Other experiments, using a wide variety of tasks and subjects, show that people are generally quite poorly calibrated. In particular, people act as though they can make much finer distinctions in their degree of uncertainty than is actually the case.

3. Overconfidence is found in most tasks; that is, people tend to overestimate how much they know.

4. The degree of overconfidence untutored assessors show is a function of the difficulty of the task. The more difficult the task, the greater the overconfidence.

5. Training can improve calibration only to a limited extent.

The second class of tasks is calibration for probabilities assigned to uncertain continuous quqantities. For example, what is the mean time between failures for this system? How much will this project cost? The assessor must report a probability density function across the possible values of such uncertain quantities. The usual method for eliciting such probability density functions is to assess a small number of fractiles of the function. The .25 fractile, for example, is that value of the uncertain quantity such that there is just a 25% chance that the true value will be smaller than the specified value. Suppose we had a person assess a large number of .25 fractiles. The assessor would be giving numbers such that, for example, "There is a 25% chance that this repair will be done in less than x_i hours" and "There is a 25% chance that Warsaw Pact personnel in Czechoslovakia number less than x_j." This person will be well calibrated if, over a large set of such estimates, just 25% of the true values turn out to be less than the x-value specified for each one. The measures of calibration used most frequently in research consider pairs of extreme fractiles. For example, experimenters assess calibration by asking whether 98% of the true values fall between an assessor's .01 and .99 fractiles.

For calibration of continuous quantities, the following results summarize the research.

1. A nearly universal bias is found: assessors' probability density functions are too narrow. For example, 20 to 50% of the true values lie outside the .01 and .99 fractiles, instead of the prescribed 2%. This bias reflects overconfidence; the assessors think they know more about the uncertain quantities than they actually do know.

2. Some data from weather forecasters suggests that they are not overconfident in this task. But it is unclear whether this is due to training, experience, special instructions, or the specific uncertain quantities they deal with (e.g., tomorrow's high temperature).

3. A few studies have indicated that, with practice, people can learn to become somewhat better calibrated.

Implications

Since assessed probabilities are central to a wide variety of decision problems (e.g., making intelligence estimates, assessing system reliability, projecting costs, deciding whether to acquire more information), the question of whether such probabilities are calibrated has far-reaching importance. Almost all decision analyses involve probability assessments. If these assessments are in error, the finest analysis relying on them may be faulty. The bias towards overconfidence reported here is widespread and well documented. What is not so well established is whether, and how, this bias can be overcome through training. The superior performance of weather forecasters is encouraging. These people have been using probabilities in their forecasts on a daily basis for several years; one might assume that this experience accounts for their excellence. Further research is needed to document just how much training, with what kind of feedback, is most efficient for improving assessors' calibration. Such research is crucial to developing a viable decision analysis technology. It also helps tell us how much faith to put in the probability assessments and decisions of untrained decision makers working without the benefit of decision aids.

Title: Lichtenstein, S. & Fischhoff, B. The effects of gender and instructions on calibration (Tech. Rep. PTR-1092-81-4). Perceptronics, Inc., July 1981.

Summary

Overview

One way that people can express their confidence in the accuracy of their own knowledge is to use probabilities (e.g., the probability that event A will occur--or that intelligence report B is true--is .75). One measure of the adequacy of probability assessments is called calibration. A set of probability assessments are well calibrated if, in the long run, the proportion of events that occur or statements that are true is equal to the assessed probability. Thus, for example, your assessments of .75 are well calibrated if just 75% of the events in question occur. The research project under which the present paper was written has as its goal to explore the psychology of confidence as expressed via probabilities.

Background

A large research literature exists on the calibration of probabilities. However, most of the research has employed naive participants who have recieved only very brief instructions concerning probability. The present report compares the calibration of participants given only the usual brief instructions with the calibration of those who were presented with lengthy instructions that more fully explained probability and calibration. In addition, the present report explores one possible cultural source of differences in confidence, gender. If it is true that males in our culture are socialized to be confident whereas females are trained to be modest, or even deprecatory, about their abilities, one might expect that females would be less confident when assessing probabilities.

Approach

The task was to decide, for each of 200 general-knowledge questions, which of two possible answers was correct (e.g., "The spleen's function is to filter [a] blood, [b] lymph") and to assess the probability that the chosen answer was indeed the correct one. About half of the 34 male and 37 female subjects were given short instructions; the others were given long instructions.

Findings and Implications

There was no effect on calibration or confidence due to instructions. This finding is consistent with previous research suggesting that overconfidence is more related to cognitive difficulties than to unfamiliarity with the response scale.

In addition, males and females did not differ with respect to calibration or confidence.

Title: Fischhoff, B. & Bar-Hillel, M. Diagnosticity and the base-rate effect (PTR-1092-81-11). Perceptronics, Inc., November 1981.

Summary

When making predictions, one is often called upon to combine two kinds of information: (a) base-rate information, describing what usually happens in such situations; (b) individuating information, describing the unique features of the particular case in point. These contrasting kinds of information may be found in such varied problems as predicting whether a particular individual is likely to commit a violent act or predicting whether a particular individual will respond positively to a charity appeal. In the first of these examples, the base rate might be the prevalence of violent individuals in the general population and the individuating information might be a detailed personality profile; in the second case, the base rate might be the population donor rate and the individuating information the solicitor's first impression.

The degree to which one relies on these two kinds of information should reflect the relative quality of the information that they provide. A lively controversy in the research literature has considered the extent to which people intuitively follow this normative judgmental rule. The present studies look at predictions made for tasks in which base-rate information is perfect and individuating information is given by no more than a thumbnail description of the target individuals. These descriptions vary to the extent that they allow the use of a judgmental rule known as "representativeness," according to which an individual is judged to belong to a category (e.g., violent, non-violent) whose stereotype he or she resembles. Representativeness was measured directly in a separate study.

It was found that people ignore base rates whenever they are able to rely on representativeness. When representativeness fails to provide a guide, they

attend to base rates; however, they have little confidence in their predictions. Confidence seems to be directly related to the ability to use representativeness--although there is a slight tendency to show reduced confidence when representativeness leads one to predict an event with a low base rate.

In addition to clarifying the conflicting results in the literature and to offering several converging techniques for measuring representativeness, the present study has a clear message for those interested in improving judgment: judges need to receive direct training in how to evaluate information.

Title: Fischhoff, B. & MacGregor, D. Subjective confidence in forecasts (Tech. Rep. PTR-1092-81-12). Perceptronics, Inc., December 1981.

Archival Publication: Fischhoff, B. & MacGregor, D. Subjective confidence in forecasts. Journal of Forecasting, 1982, 1, 155-172.

Summary

Forecasts have little value to decisions makers unless it is known how much confidence to place in them. Those expressions of confidence have, in turn, little value unless forecasters are able to assess the limits of their own knowledge accurately.

Previous research has shown very robust patterns in the judgments of individuals who have not received special training in confidence assessment: Knowledge generally increases as confidence increases. However, it increases too swiftly, with a doubling of confidence being associated with perhaps a 50% increase in knowledge. With all but the easiest of tasks, people tend to be overconfident regarding how much they know.

These results have typically been derived from studies of judgments of general knowledge. The present study found that they also pertained to confidence in forecasts. Indeed, the confidence-knowledge curves observed here were strikingly similar to those observed previously. The only deviation was the discovery that a substantial minority of judges never expressed complete confidence in any of their forecasts. These individuals also proved to be better assessors of the extent of their own knowledge.

Apparently confidence in forecasts is determined by processes similar to those that determine confidence in general knowledge. Decision makers can use forecasters' assessments in a relative sense, in order to predict when they are more and less likely to be correct. However, they should be hesitant to take confidence assessments literally. Someone is more likely to be right when he or she is "certain" than when he or she is "fairly confident;" but there is no guarantee that the certain forecast will come true.

Title: Kadane, J. & Lichtenstein, S. A subjectivist view of calibration (Tech. Rep. PTR-1092-82-7). Perceptronics, Inc., July 1982.

Summary

Calibration concerns the relationship between subjective probabil'ties and the long-run frequencies of events. Theorems from the statistical and probability literature are reviewed to discover the conditions under which a coherent Bayesian expects to be calibrated. If the probability assessor knows the outcomes of all previous events when making each assessment, calibration is always expected. However, when such outcome feedback is lacking, the assessor expects to be well calibrated on an exchangeable set of events if and only if all the events in question are viewed as independent. Although this strong condition has not been tested in previous research, we speculate that the past findings of pervasive overconfidence are not invalid. Although experimental studies of calibration hold promise for the development of cognitive theories of confidence, their value for the practice of probability assessment seems more limited. Efforts to train probability assessors to be calibrated may be misplaced.

Title: Fischhoff, B. & Beyth-Marom, R. Hypothesis testing from a Bayesian perspective (Tech. Rep. PTR-1092-82-6). Perceptronics, Inc., July 1982.

Archival Publication: Fischhoff, B. & Beyth-Marom, R. Hypothesis evaluation from a Bayesian perspective. Psychological Review, 1983, 90, 239-260.

Summary

Bayesian inference provides a general framework for evaluating hypotheses. It is a normative method, in the sense of prescribing how hypotheses should be evaluated. However, it may also be used descriptively, by characterizing people's actual hypothesis evaluation behavior in terms of its consistency with or departures from the model. Such a characterization may facilitate the development of psychological accounts of how that behavior is produced (i.e., as the result of failed attempts to act in a Bayesian fashion or as the result of attempts to process information in non-Bayesian ways).

This essay exploits the descriptive potential of Bayesian inference. First, it identifies a set of logically possible forms of non-Bayesian behavior listed according to the task in which the problem arises: hypothesis formulation, assessing component probabilities, assessing prior odds, assessing likelihood ratio, aggregation, information search or action. For example, one potential failure in hypothesis formulation is for the hypothesis to be untestable as a result of being ambiguous and/or complex.

The essay then proceeds to apply this framework in a selective review of existing research in a variety of areas. It does so with a triple purpose: (a) to illustrate the different biases, (b) to identify which of the set of all possible biases have in fact been observed and documented, and (c) to show how the entire hypothesis evaluation process must be considered when characterizing any one individual behavior.

Several conclusions emerge from the analysis: (a) In some situations, several phenomena that have previously been thought of as distinct may be usefully viewed as special cases of the same behavior. The neglect of the alternative hypothesis in assessing the likelihood ratio is one such example of a simple phenomenon that has been differently labeled in different contexts (pseudodiagnosticity, inertia, cold readings).

(b) In other situations, previous investigations have conferred a common label to several distinct phenomena (for example, "confirmation bias" has been applied to acts that might be better characterized as "failure to ask potentially falsifying questions," "asking non-diagnostic questions," "mistaking affirmation for confirmation," etc.).

(c) It calls into question a number of attributions of judgmental bias, suggesting that in some cases the bias is different than what has previously been claimed, whereas in others, there may be no bias at all.

Title: Fischhoff, B., MacGregor, D. & Lichtenstein, S. Categorical Confidence (Tech. Rep. PTR-1092-82-7). Perceptronics, Inc., July 1982.

Summary

People tend to be inadequately sensitive to the extent of their own knowledge when asked to assess the probability that each of their answers to a set of questions is correct. This insensitivity typically emerges as overconfidence. That is, their assessments are typically too high compared to the portion of items they get right. Few prescriptions have proven effective against this problem. Those that have worked might be thought of as directive in character. Rather than improving subjects' feelings for how much they know, they may have suggested to subjects how their probability assessments should be changed. These successful manipulations include giving feedback and requiring subjects to provide reasons contradicting their chosen answers. The present study attempted to improve the appropriateness of confidence with a seemingly undirective tack. Subjects were asked to sort items into a specified number of piles according to their confidence in the correctness of their answers. Subsequently, they assessed, for each pile, the probability that each of the items in it was correct. Even though this procedure differed from its predecessors in many respects, performance here was indistinguishable from that observed elsewhere. Though small pockets of improvement were noted, confidence was largely resistant to this manipulation. Some implications of these results for attempts to study confidence and eliminate overconfidence are discussed.

Title: Beyth-Marom, R. & Fischhoff, B. Diagnosticity and pseudodiagnosticity (Tech. Rep. PFTR-1092-83-1A). Perceptronics, Inc., January 1983.

Archival Publication: Beyth-Marom, R. & Fischhoff, B. Diagnosticity and pseudo-diagnosticity. Journal of Personality and Social Psychology, in press.

Summary

Five experiments contrasted subjects' intuitive evaluation of data for hypothesis testing with the Bayesian concept of diagnosticity. According to that normative model, the impact of a datum, D, relative to a pair of hypotheses, H and $\bar{H}$, is captured by its likelihood ratio, equal to $P(D/H)/P(D/\bar{H})$. The studies found that when subjects were asked to test the validity of H, only half expressed an interest in $P(D/\bar{H})$. That proportion increased when they were asked to determine whether H or $\bar{H}$ is true. That proportion decreased when the instructions more forcefully encouraged subjects to solicit only pertinent information. Thus, subjects generally had a strong interest only in the conditional probability that mentions the hypothesis (or hypotheses) that they are explicitly asked to test. When, however, they were presented with both components of the likelihood ratio, most subjects revealed a qualitative understanding of their meaning vis-a-vis hypothesis testing. These results are discussed in terms of the kinds of understandings that people might have for statistical principles.

Title: Beyth-Marom, R. & Lichtenstein, S. Accuracy of confidence assessments: Does number of alternatives make a difference? (Tech Rep. PFTR-1092-83-1B). Perceptronics, Inc., January 1983.

Summary

The accuracy (calibration) of probabilistic confidence assessments was studied using one-alternative and two-alternative general knowledge items. Subjects were first given either 10 one-alternative items or 10 two-alternative items and asked to perform the usual calibration task: (a) choose the correct alternative (for the two-alternative task) or state whether the given alternative was true or false (for the one-alternative task) and (b) assess the probability that the choice or answer was correct. Secondly, for a new set of 10 items subjects were asked to list supporting and contradicting reasons either to the given alternative (one-alternative task) or to a prespecified alternative (two-alternative task). Finally, they again performed the usual calibration task, this time for the items they had just given reasons to. Subjects were better calibrated and less overconfident on the two-alternative task than on the one-alternative task. A higher proportion of subjects avoided using 1.0 responses in the two-alternative task than in the one-alternative task. Non-users of the 1.0 response were better calibrated than users. The elicitation of reasons did not improve calibration.